Modified K-Means for Document Clustering System

Tin Thu Zar Win, and Moe Moe Aye

***Abstract*— In today's era of World Wide Web, there is a tremendous proliferation in the amount of digitized text documents. As there is huge collection of documents on the web, there is a need of grouping the set of documents into clusters. Document clustering plays an important role in effectively navigating and organizing the documents. K-Means clustering algorithm is the most commonly document clustering algorithm because it can be easily implemented and is the most efficient one in terms of execution times. The major problem with this algorithm is that it is quite sensitive to selection of initial cluster centroids. The algorithm takes the initial cluster center arbitrarily so it does not always promise good clustering results. If the initial centroids are incorrectly determined, the remaining data points with the same similarity scores may fall into the different clusters instead of the same cluster. To overcome this problem, modified K-Means approach is proposed to improve the quality of clustering in this paper. Unlike the traditional K-Means clustering, the proposed K-Means method can generate the most compact and stable clustering results based on maximum distance initial centroids points instead of random initial centroid points. Moreover, the similar data points are clustered based on maximum probability distribution of data points. Therefore, the proposed method is more effective and converges to more accurate clusters than original K-Means clustering method. In this paper, experimental results are presented in F-measure using 20-News Group standard dataset.**

***Keywords*— Document clustering, F-measure, Initial centroid, K-Means**

I. Introduction

Due to the rapid advancement of smart technologies in World Wide Web, the volume of digitized text documents has been increasing rapidly. Clustering plays an important role for organizing such massive document collection returned by search engines into meaningful clusters. Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. The purpose of clustering is to introduce an order in a collection of documents by grouping or classification. Document clustering has been studied intensively because of its wide applicability in areas such as web mining, search engines, information retrieval, and topological analysis [1].

The fast and high-quality document clustering algorithms play an important role in helping users to effectively navigate, summarize, and organize the information. There are various challenges and requirements of clustering algorithm to increase the quality of clusters. Various clustering algorithms are available to cluster text documents. Data clustering can be broadly categorized into hierarchical methods, partition methods, fuzzy clustering methods, hard clustering methods and model-based methods. Basically, the two main approaches to document clustering are Hierarchical Clustering method and Partitioning Clustering method [2].

Hierarchical Clustering method often defines as a better quality clustering approach, but it is limited because of its quadratic time complexity and also it does not contain any provision for the reallocation of entities. Partitional clustering algorithms partition the data set into a particular number of clusters and then evaluate them on the basis of a criterion. In partitioning clustering method, K-Means clustering algorithm is one of the most widely used partition based clustering algorithm that arrange the documents in order such that a document is close to its related document on the basis of similarity measure. It is suited for clustering a large document dataset due to its linear time complexity. Moreover, it is simple to understand and computationally efficient. However, it is sensitive to random selection of initial centroids. Different initial cluster centers often leads to different clustering and thus provide unstable clustering results. So, the better choice is to place them as much as possible far away from each other. In addition to selection of initial centroid, the output of the clustering algorithm also depends on the clustering techniques [3]. Several improved methods have been proposed to achieve optimized K-Means algorithm. In this paper, the proposed algorithm is presented by modifying simple K-Means algorithm.

The rest of the paper is organized as follows: the details of background theory are discussed in Section II. The proposed model of this system is briefly explained in Section III. The experimental results are presented in Section IV. The paper is concluded in section V.

II. Background Theory

A. Clustering

Cluster analysis is one of the major data analysis methods widely used for many practical applications in emerging areas. Clustering is the process of finding groups of objects such that the objects in a group will be similar to one another and different from the objects in other groups. A good clustering method will

Manuscript received Nov. 1, 2016.

Tin Thu Zar Win
Department of Computer Engineering and Information Technology, Mandalay Technological University, Myanmar, (e-mail: zarzar84mtu@gmail.com).

Moe Moe Aye
Department of Computer Engineering and Information Technology, Mandalay Technological University, Myanmar, (e-mail: moeaye255@gmail.com).

produce high quality clusters with high intra-cluster similarity and low inter-cluster similarity. The quality of a clustering result depends on both the similarity measure used by the method and its implementation and also by its ability to discover some or all of the hidden patterns [4].

B. K-Means Clustering

K-Means clustering is one of the unsupervised computational methods used to group similar objects in to smaller partitions called clusters so that similar objects are grouped together. The algorithm aims to minimize the within cluster variance and maximize the intra cluster's variance. The K-Means clustering algorithm is one of the simplest clustering algorithms in which the number of clusters to be grouped is fixed a priori by the user. The algorithm proceeds by randomly defining *k* centroids and assigning a document to the cluster that has the nearest centroid to the document [5]. Then, for every data point, the minimum distance is determined and that point is assigned to the closest cluster. This step is called cluster assignment, and is repeated until all of the data points have been assigned to one of the clusters. Finally, the mean for each cluster is calculated based on the accumulated values of points in each cluster and the number of points in that cluster. Those means are then assigned as new cluster centroids, and the process of finding distances between each point and the new centroids is repeated, where points are re-assigned to the new closest clusters. The process iterates for a fixed number of times, or until points in each cluster stop moving across to different clusters. This is called convergence [6]. The flowchart of the K-Means algorithm is as shown in Fig. 1.

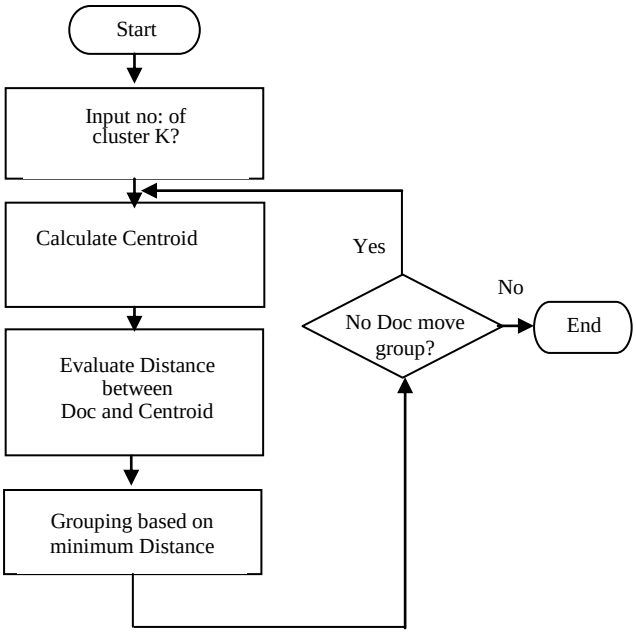

Fig. 1 Flow Chart of K-Means Algorithm [8]

Euclidean metric is considered as it is one of the widely used distance metrics incorporated with K-Means clustering and one that is easy to implement. Also it results in a best solution. Euclidean distance is given in equation 1:

$$\text{dist}(P, C) = \sqrt{\sum_{i=1}^{n} (P_i - C_i)^2} \quad \text{_____ (1)}$$

Where P is the data point, C is the cluster center, and n is the number of features.

C. Characteristics of K-Means Clustering

The characteristics of K-Means clustering are described as follow [7]:

1. The quality of the cluster can be measured by the variation within cluster, which is in terms of sum of squared error among all objects in that cluster.
2. The K-Means method is not guaranteed to converge to the global optimum and often terminates at a local optimum. The results may depend on the initial random selection of cluster centers.
3. The time complexity of the K-Means algorithm is O(nkt), where n is the total number of elements or data points in the dataset D, k is the number of clusters, and t is the number of iterations.
4. Generally, k << n and t << n, so the method is relatively scalable and efficient in processing large data sets.
5. To obtain the good results, it is common to run the K-Means algorithm multiple times with the different initial cluster centers.

D. K-Means Clustering Algorithm

The K-Means algorithm is composed of the following steps [8]:

Input: Number of desired clusters, k, and a database D={d1,d2,…dn} containing n data objects.

Output: A set of k clusters

Steps:

1. Here we have to select randomly k data objects from dataset D as initial cluster centers.
2. Repeat;
3. Then calculate the distance between each data object di (1 <= i<=n) and all k cluster centers cj(1<=j<=k) and assign data object di to the nearest (closest) cluster.
4. For each cluster j, recalculate the cluster center.
5. Until no changing in the center of clusters.

III. PROPOSED CLUSTERING MODEL

The idea of the proposed model is to cluster the documents by modifying the initial centroid selection method and clustering nature of similar documents for K-Means algorithm. Instead of randomness initial centroid of original K-Means, it starts by finding the maximum distance objects as initial centroids. The right choice of initial centroid is very important for K-Means clustering process. In addition, the new probability distribution equation which is described in equation 2 is used for grouping the remaining documents in the dataset.

$$P(Di, Cj) = \sum_{i=1}^{k} P(Dwi) * P(wi \mid Cj) \quad \text{_____ (2)}$$

$$\text{Where, } P(Dwi) = \frac{\text{no : of wi count in selected Doc}}{\text{Total word count in selected Doc}}$$

$$P(wi \mid Cj) = \frac{\text{no : of wi count in cluster}}{\text{Total word count in cluster}}$$

In the proposed model, documents which need to be clustered are loaded from the database. After that, those documents are pre-processed using stop word removal and stemming algorithm. After pre-processing documents, the modified K-Means algorithm is performed. The flow chart of the proposed model is shown in Fig. 2.

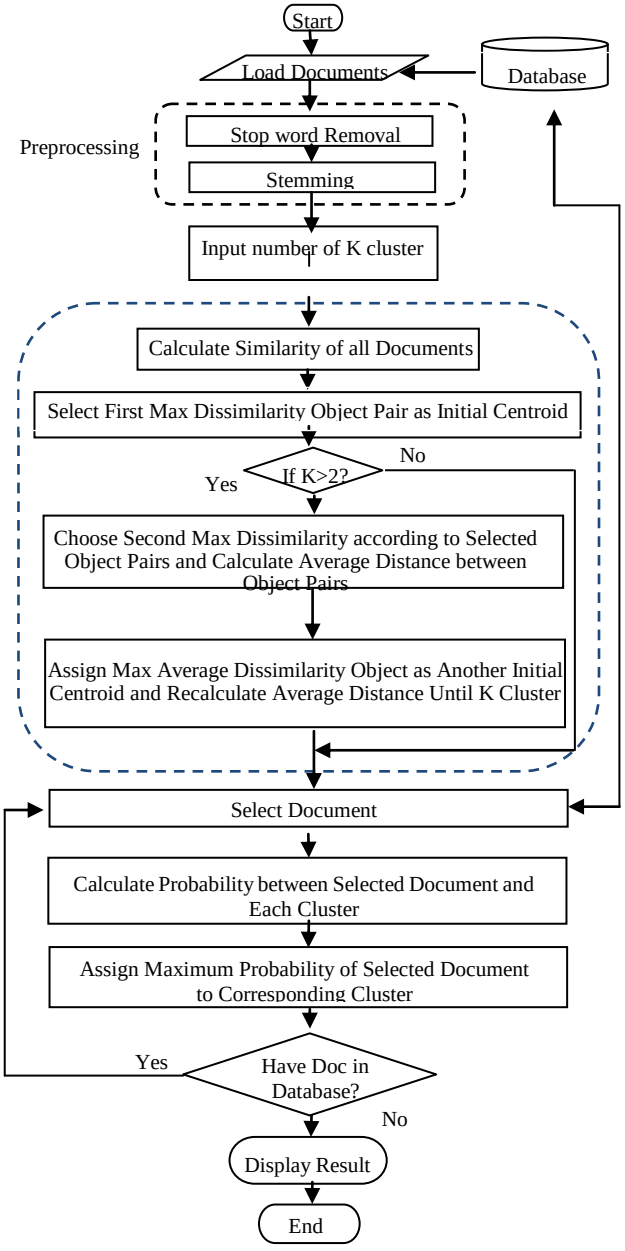

Fig. 2 The procedures of the proposed model

In modified K-Means algorithm, two main processes; initial centroid selection and clustering similar documents with probability distribution are performed. In initial centroid selection process, the initial centroids are determined based on maximum distance object pair. If the number of k cluster is greater than 2, the other initial points are determined based on average distance between furthest data points and the maximum average distance object are assigned as the other initial centroid. So, average maximum object is recalculated until the number of k-cluster. Selecting maximum distance centroids approach is the main part of K-Means clustering because this can affect the accuracy of clustering.

In second process, the remaining documents are grouped depending on the result of initial centroid selection process. After the initial centroids are selected systematically, the relationship between the selected document and the previous cluster is calculated by using new probability distribution equation which is shown in equation 2. So, the probability of each document is needed to calculate for clustering to corresponding cluster. According to probability result, the document which has maximum probability value is assigned to its' related cluster. So, the probability of all documents is recalculated as an iterative process until no more documents in dataset.

A. Modified K-Means Clustering Algorithm

The procedure of the modified K-Means algorithm is as follows:

1. Accept k-input number.
2. Calculate similarity metric of all documents from dataset.
3. Select first maximum distance document pair as initial centroids
4. If k>2, then
   (i) Choose second maximum distance according to selected document pairs and calculate the average distance between document pairs.
   (ii) Assign maximum average distance document as another initial centroid.
   Else go to step (6).
5. Repeat step (4) until to K cluster.
6. Select one document from dataset; calculate probability between selected document and each cluster, and then assign maximum probability of selected document to its corresponding cluster.
7. Repeat step (6) until no more document in dataset.

B. Detail Explanation of K-Means and Modified K-Means

It is assumed that there are five documents, D_0 to D_4 and number of cluster, k=2. Five documents are as follows:

D_0= information retrieval system extract information database
D_1= wireless sensor network technology important sensor node
D_2= database information retrieval system
D_3= relevance information important information retrieval
D_4= sensor node information network

First, all documents are converted into the data representation matrix and counted the term frequency. This processing step is

necessary for both simple K-Means and modified K-Means algorithm. It is supposed that these five documents are passed the pre-processing stage. The data matrix is shown in Table I.

TABLE I: DATA REPRESENTATION MATRIX

Terms	Documents				
	D_0	D_1	D_2	D_3	D_4
information	2	0	1	2	1
retrieval	1	0	1	1	0
system	1	0	1	0	0
extract	1	0	0	0	0
database	1	0	1	0	0
relevance	0	0	0	1	0
important	0	1	0	1	0
wireless	0	1	0	0	0
sensor	0	2	0	0	1
network	0	1	0	0	1
technology	0	1	0	0	0
node	0	1	0	0	1

1. By using K-Means method,

According to K-Means, the algorithm starts by randomly defining *k* centroids from among these five documents. In this example, it is supposed that two documents (D_0 and D_3) are randomly chosen as initial centroids and these two documents are assigned into cluster C_1 and cluster C_2 as follow:

Cluster 1	Cluster 2
D_0	D_3

Secondly, the distance measures between documents and centroids are calculated by using Euclidean distance. The results of distance measure are shown in Table II.

TABLE II: DISTANCE MATRIX USING EUCLIDEAN DISTANCE MEASURE

Document	Centroid 1	Centroid2
D_0	0	2.236
D_1	4.123	3.742
D_2	1.414	2.236
D_3	2.236	0
D_4	2.828	2.646

Then, for every data point, the minimum distance is determined and that point is assigned to the closest cluster. This step is called cluster assignment, and is repeated until all of the data points have been assigned to one of the clusters. The result of documents cluster is as follow:

Cluster 1	Cluster 2
D_0 ,D_2	D_1 ,D_3 ,D_4

After that, the new mean for each cluster is calculated based on the accumulated values of points in each cluster. Those means are then assigned as new cluster centroids. So, the new centroids are as follow:

Centroid C_1 = (3/2, 1, 1, 1/2, 1, 0, 0, 0, 0, 0, 0, 0)

Centroid C_2 = (1, 1/3, 0, 0, 0, 1/3, 2/3, 1/3, 1, 2/3, 1/3, 2/3)

The process of finding distances between each point and the new centroids is repeated, where points are re-assigned to the new closest clusters. The results are shown in Table III.

TABLE III: DISTANCE MATRIX USING EUCLIDEAN DISTANCE MEASURE

Document	Centroid 1	Centroid2
D_0	0.707	2.667
D_1	3.808	1.856
D2	0.707	2.601
D_3	2.121	2.028
D_4	2.549	1.054

The process iterates until no more change cluster result in each cluster. Now, the result is no more change as previous step. So, the final resultant clusters are as follow:

Final Resultant Clusters

Cluster 1	Cluster 2
D_0 ,D_2	D_1 , D_3 ,D_4

2. By using Modified K-Means method,

In modified K-Means method, constructing the data representation matrix and preprocessing steps are performed. The result of data matrix is also shown in Table I.

After constructing the data representation matrix, Euclidean distance is used to measure the similarity between pair of documents. The resulting similarity matrix is shown in Table IV.

D_0= information retrieval system extract information database
D_1= wireless sensor network technology important sensor node
D_2= database information retrieval system
D_3= relevance information important information retrieval
D_4= sensor node information network

TABLE IV: SIMILARITY MATRIX USING EUCLIDEAN DISTANCE MEASURE

Doc	D_0	D_1	D_2	D_3	D_4
D_0	0	4.123	1.414	2.236	2.828
D_1	4.123	0	3.606	3.742	2.236
D_2	1.414	3.606	0	2.236	2.449
D_3	2.236	3.742	2.236	0	2.646
D_4	2.828	2.236	2.449	2.646	0

Based on the similarity matrix, the maximum distance is 4.123. Therefore, the maximum distance document pair (D_0 and D_1) is selected as first initial centroid. If k is not greater than 2, another maximum documents pairs are not selected as initial centroids. So, the two initial centroids are assigned to its cluster as follow:

Cluster1	Cluster2
D_0	D_1

After that, the remaining documents in the dataset are clustered by using the innovative idea for distribution of documents. Instead of repeat iteration of final cluster result of original K-Means, the proposed algorithm merges by finding the maximum probability distribution of objects to its corresponding cluster. Document D_2 is selected and the probability of D_2 is calculated depending on cluster C_1 and cluster C_2 by using equation 2. The calculation of probability results are shown as follow:

Doc D_2= database information retrieval system
P (database) =1/4, P (information) = 1/4, P (retrieval) = 1/4,
P (system) = ¼

$P(D_2, C_1) = ?$

$P(database \mid C_1) = 1/6$
$P(information \mid C_1) = 2/6$
$P(retrieval \mid C_1) = 1/6$
$P(system \mid C_1) = 1/6$

$P(D_2, C_1) = (1/6 * 1/4) + (2/6 * 1/4) + (1/6 * 1/4) + (1/6 * 1/4)$
$= 0.208$

$P(D_2, C_2) = ?$

$P(database \mid C_2) = 0$
$P(information \mid C_2) = 0$
$P(retrieval \mid C_2) = 0$
$P(system \mid C_2) = 0$

$P(D_2, C_2) = 0$

According to the result, Document D_2 has maximum probability value, 0.208 depending on cluster C_1. So, D_2 is merged to cluster C_1 as follow:

Cluster1	Cluster2
D_0 , D_2	D_1

Then, document D_3 is selected and the probability of document D_3 is calculated depending on cluster C_1 and C_2. Document D_3 has maximum probability value, 0.16 depend on Cluster C_1. But it has probability value, 0.029 depend on Cluster C_2. So, it is merged to cluster C_1 .The calculation steps are described as follow:

D_3= relevance information important information retrieval
P (relevance) =1/5, P (information) = 2/5, P (important) = 1/5,
P (retrieval) = 1/5

$P(D_3, C_1) = ?$

$P(relevance \mid C_1) = 0$
$P(information \mid C_1) = 3/10$
$P(important \mid C_1) = 0$
$P(retrieval \mid C_1) = 2/10$

$P(D_3, C_1) = 0 + (3/10 * 2/5) + 0 + (2/10 * 1/5)$
$= 0.16$

$P(D_3, C_2) = ?$

$P(relevance \mid C_2) = 0$
$P(information \mid C_2) = 0$
$P(important \mid C_2) = 1/7$
$P(retrieval \mid C_2) = 0$

$P(D_3, C_2) = (1/7 * 1/5)$
$= 0.029$

Cluster1	Cluster2
D_0 , D_2 , D_3	D_1

The probability of Doc D_4 is calculated as previous steps and the calculation steps are skipped. When all documents in the database are merged to its corresponding cluster, the process is terminated. The final resultant clusters are shown as follow:

Final Resultant Clusters

Cluster 1	Cluster 2
D_0 , D_2 , D_3	D_1 , D_4

C. Performance Comparison between K-Means and Modified K-Means

In Modified K-Means method, the document D_1 is more closely related with D_4 and document D_0, D_2 are more closely related with D_3. The algorithm merges correctly the documents to its cluster. However, the existing K-Means method merges D_1 with other documents (D_3 and D_4). So, the existing K-Means method cannot merge the cluster efficiently. According to the results, it is clearly concluded that modified K-Means method can give the better cluster quality than existing K-Means method.

IV. EXPERIMENTAL RESULTS

In this section, the experimental results are presented. The experiment has been carried out only for initial centroid selection process of the proposed method. The remaining process of the proposed method, clustering similar documents with probability distribution, is still in progress. Firstly, the maximum distance objects are defined as initial centroids. The remaining procedure is almost similar to the original K-Means algorithm except that the initial centroids are computed systematically. For testing accuracy and efficiency of simple K-Means and modified K-Means algorithm, 20-newsgroups dataset is chosen. It is a collection of news articles collected from 20 different sources. The 100 documents are randomly selected from ten categories and developed the dataset 20ns which consists of 1000 documents.

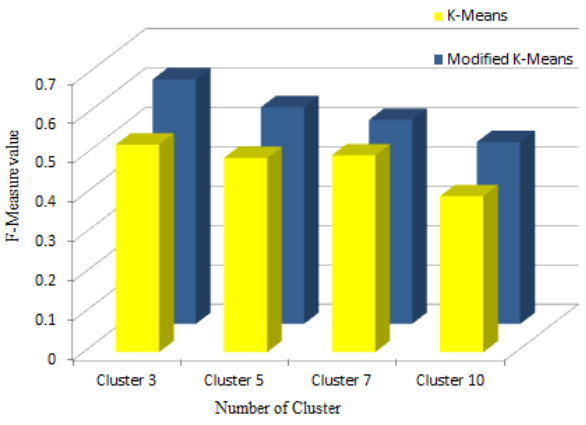

Fig. 3 Comparison of F-Measure between K-Means and modified K-Means on 20_ NewsGroup datasets

In Fig. 3, the dataset 20ns is tested in F-measure for various numbers of clusters. It can be clearly seen that traditional K-Means method is less than in F-measure than the modified K-Means method for various number of cluster. It may be noted that the proposed method outperforms the traditional K-Means in terms of cluster quality.

V. Conclusion

In this paper, an effective K-Means approach for document clustering system is presented. According to the calculation, which is considered both for initial centroid selection process and clustering similar documents with probability distribution process, the proposed centroid selection process selects the most dissimilar documents as initial centroid points. Therefore, this system can merge clusters together efficiently and can also give the better cluster quality. In addition, the remaining documents are clustered in accordance with the probability distribution of documents that provides more accurate cluster result. Because of the systematic selection of initial centroid and probability distribution of documents processes, the proposed method produces significantly better clustering solutions. According to the experiments, which is considered only for initial centroid selection process, it is observed that the initial centroid selection approach outperforms the traditional K-Means algorithm in term of cluster quality and can be applied for unsupervised clustering of various datasets. Moreover, it is helpful in selecting significant centers for K-Means clustering. How to really perfect the proposed model requires further research. In future, the proposed model requires further research to give perfect clustering approach by experimenting with probability distribution of documents. Then, this system is experimented with a number of benchmark datasets such as *tr12*, *tr23*, *tr45* are derived from TREC-5, TREC-6 and TREC-7 text collections, Reuters-21578 and 20-NewsGroup to calculate the cluster accuracy using F-measure.

Acknowledgment

The author would like to thank Dr. Nang Aye Aye Htwe, Associate Professor, Head of Department of Computer Engineering and Information Technology for her encouragement and kindly advice. The author would like to especially express her deep appreciation to her supervisor, Dr. Moe Moe Aye, Associate Professor, Department of Computer Engineering and Information Technology, Mandalay Technological University for her close supervision, helpful advice, encouragement and invaluable guidance. The author would also thank to her parents, all her friends and all the teachers who taught her throughout the whole life.

References

[1] M. konchady, "Text Mining Application Programming," Programming Series, Charles River Media, 2006.

[2] Huang, "Similarity Measures for Text Document Clustering," Proc. of the 6th New Zealand Computer Science Research Student Conference NZCSRSC, pp. 49-56, 2008.

[3] H. Neepa, et al, "Distributed Document Clustering Using K-Means," International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Issue 11, November 2014.

[4] P. Larma, "Clustering System Based on Text Mining Using The K-Means Algorithm," Turku University of Applied Sciences Thesis, Information Technology, December 2013.

[5] M. Steinbach, G. Karypis, V. Kumar, "A comparison of document clustering techniques," proc. KDD Workshop on Text Mining, 1-20, 2000.

[6] Swapnali Ware, N. A. Dhawas, "Web Document Clustering Using KEA-Means Algorithm," Int. J. Computer Technology & Applications, Vol 3 (5), 1720 -1725, 2011.

[7] R. Sharma, M. Afshar Alam, "K-Means Clustering in Spacial Data Mining using Weka Interface," International Conference on Advances in Communication and Computing Technologies (ICACACT) 2012, Proceedings published by International Journal of Computer Applications (IJCA).

[8] Teknomo, K., "K-Means Clustering Tutorials," (2007).

Tin Thu Zar Win is now a Ph.D student of Mandalay Technological University, Mandalay, Myanmar. She was born in Yangon, The Republic of Union of Myanmar. Her date of birth is 1 June, 1984. She got her Bachelor of Engineering (B.E) degree with Information Technology from Technological University (Thanlyin), Myanmar in 2006. She obtained her Master of Engineering (ME) degree with Information Technology from Mandalay Technological University, Myanmar in 2014.

She has over nine years working experience in research and in teaching. Now, she is also doing her Ph.D research in Computer Engineering and Information Technology at Mandalay Technological University. She participated in International Journal of Scientific Engineering and Technology Research by writing an article about "*Applying Mobile Agent in Criminal Information Retrieval System,"* in 2014.